

Assessment of Pooled Association Tests for Rare Genetic Variants within a Unified Framework

Andriy Derkach*, J.F. Lawless[†], Lei Sun[‡]

*PhD. Candidate, Department of Statistics, University of Toronto, 100 St. George Street, Toronto, ON M5S 3G3 (E-mail: derkach@utstat.toronto.edu).

[†]Distinguished Professor Emeritus, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1 and Adjunct Professor, Biostatistics Division, Dalla Lana School of Public Health, 155 College Street, Toronto ON M5T 3M7 (Email: jlawless@math.uwaterloo.ca).

[‡]Associate Professor, Biostatistics Division and Department of Statistics, University of Toronto, 155 College Street, Toronto, ON M5T 3M7, Canada (Email: sun@utstat.toronto.edu).

Abstract

Numerous recent papers have proposed pooled testing strategies to assess the association between a group of rare genetic variants and complex human traits. We consider a unified framework that addresses key issues in pooled testing, including the use of weights assigned to particular variants and the directions of genetic effects. We categorize methods into two classes: linear statistics sensitive to specific directional alternatives and “omnibus” quadratic statistics, which have reasonable power across a wide range of alternatives. The powers of the statistics are related to non-centrality parameters associated with normal approximations. The power for a quadratic statistic that we consider depends only on the explained variation for a group of variants. In contrast, the power for linear statistics depends on the variants’ frequencies, directions of their effects and weights, if used. The complex relationship among these factors suggests that, even if rarer variants tend to have larger genetic effects, the common strategy that chooses weights inversely proportional to a rare variant’s frequency can adversely affect power. Further, even if the causal variant effects all have the same direction, quadratic statistics can outperform linear statistics unless the proportion of causal variants in the group is sufficiently high. We also show that recent methods that use data-driven weights to maximize linear statistics are operationally similar to quadratic statistics. Our framework can deal with categorical and quantitative traits, trait-dependent selection of individuals in a study, and it extends to include stratification or adjustment for covariates. We discuss the performance of tests based on analytical results, simulation studies and applications to GAW17 data. We conclude that considerable attention should be given to background information that can guide the selection of SNPs for pooled testing.

Keywords: linear statistics; quadratic statistics; score tests; weighting; power; next generation sequencing.

1 Introduction

Genome-wide association studies (GWAS) have identified numerous genetic variants (single nucleotide polymorphisms, or SNPs) that are associated with complex diseases or traits (e.g. Manolio et al. (2008), Hindorff et al. (2009)). However, because of their limited sample size such studies are effective only at identifying common variants, that is, for which the minor allele frequency (MAF) is not too small (e.g. $\text{MAF} \geq 5\%$ for sample size ~ 2000). In addition, variants that have been identified through GWAS explain only small fractions of the genetic components of disease risks or variability in traits (Manolio and Collins (2009)). There is now mounting evidence that rare variants (as represented by SNPs with small MAFs) may contribute significantly to phenotype variation but because they are rare, their discovery is more difficult (e.g. Li and Leal (2008); Bansal et al. (2010); Asimit and Zeggini (2010)). Since large-scale studies involving huge numbers of individuals might not be a viable option due to cost, heterogeneity and other concerns, attention has focused on methods that combine information across multiple rare SNPs in a genomic region, based on data generated from the next generation sequencing (NGS) technology. This area is the focus of our article.

Papers that propose pooled association testing strategies based on the combination of information across multiple SNPs include Morgenthaler and Thilly (2007), Li and Leal (2008), Madsen and Browning (2009), Bansal et al. (2010), Han and Pan (2010), Hoffmann et al. (2010), Morris and Zeggini (2010), Price et al. (2010), Yi and Zhi (2011), Neale et al. (2011), Wu et al. (2011) and Lee et al. (2012). This previous work provided various solutions but insight into settings when a method will perform well, indifferently or poorly is still limited. Recently, Lin and Tang (2011) have given a theoretical and empirical evaluation of the methods in the preceding papers, and have

proposed a new test statistic based on the maximum absolute value across a set of statistics. Lee et al. (2012) have also compared previous tests and give a new family which they term SKAT-O, for “sequence kernel association test - optimal”. Basu and Pan (2011) have conducted an extensive empirical evaluation study of these methods in the context of case-control studies.

Comparisons across many of the proceeding papers do not result in firm conclusions about which method to use when preparing to address a specific genetic association testing setting. Unfortunately, the precise prior information essential to an informed decision concerning the choice of test statistic is in any case usually unavailable. Moreover, the term “optimal”, used in connection with any of the many classes of proposed tests, has meaning in a theoretical framework where precise knowledge exists of the true underlying phenomena, but the relevance to real settings is unclear. Basu and Pan (2011) reached more specific conclusions, which we discuss later, but recommend that when prior information is absent, both linear and quadratic statistics (see below) be used.

In this paper we consider tests for genotype-phenotype association within a unified framework. Most test statistics that have been proposed can be divided into two classes: linear composite statistics which are powerful against specific association alternatives (e.g. Morgenthaler and Thilly (2007), Li and Leal (2008), Morris and Zeggini (2010), Madsen and Browning (2009) and Price et al. (2010)), and quadratic statistics that have reasonable power across a wide range of alternatives (e.g. Neale et al. (2011), Wu et al. (2011)). A feature of many of the linear statistics and of the quadratic statistics of Wu et al. (2011) and Lee et al. (2012) is the use of weights associated with individual SNPs, one rationale being that there is evidence suggesting that larger effects are associated with rarer SNPs. We study both classes of statistics theoretically and empirically, and provide new insights. We consider scenarios that

involve varying proportions of beneficial (protective), harmful (deleterious) and neutral SNPs. We show that although linear statistics can perform well if the majority of SNPs are causal and harmful (or causal and beneficial), they perform poorly relative to quadratic statistics when there are both protective and deleterious SNPs and more generally, where a substantial portion of the SNPs under consideration are neutral. We also demonstrate that the effects of using weights in linear or quadratic statistics that are inversely proportional to a variant's MAF depend on the type of statistic. Even if the assumption that rarer variants tend to have larger genetic effects is true, such weights can in some cases have an adverse effect and in others a beneficial effect. We do not specifically study statistics using adaptive weight selection; however, we show that for linear statistics, adaptive methods of weight selection without external information (e.g. Han and Pan (2010), Yi and Zhi (2011), Hoffmann et al. (2010), Lin and Tang (2011)) are operationally similar to using quadratic statistics. We also consider the question of optimality and indicate why it is in practice unachievable. Our discussion deals with all types of traits (categorical, quantitative) and allows trait - dependent selection of individuals in a study or non-independent SNPs. It also extends to include stratification or adjustment for covariates. Like Basu and Pan (2011), Lin and Tang (2011), Wu et al. (2011) and Lee et al. (2012), we focus on score statistics, which are both theoretically and computationally efficient. Our results are relatively transparent and easy to apply to practical situations.

The remainder of the paper is organized as follows. Section 2 introduces notation and the framework for testing for association between rare variants and phenotypes. We consider arbitrary traits in Section 2 and then the important case of binary traits in Section 3. Section 4 presents theoretical power calculations for quantitative traits that indicate when various methods will do well, and Section 5 gives simulation results comparing different approaches with case-control settings. Section 6 examines data

from GAW17 data provided by the 1000 Genomes Project (Almasy et al. (2011), 1000 Genomes Project Consortium (2010)). Section 7 discusses extensions to deal with covariates. Section 8 concludes with some recommendations for pooled testing.

2 General Traits

We assume that a group of J SNPs labelled $j = 1, \dots, J$ and a trait Y are under consideration. The objective is to consider whether there is association between Y and one or more of the SNPs; we do this by formally testing the hypothesis of no association. Let A_j and B_j represent the rare and common alleles for SNP j , respectively. For a set of n unrelated individuals, let Y_i be the measured trait and X_{ij} denote the genotype for the i^{th} ($i = 1, \dots, n$) person's j^{th} ($j = 1, \dots, J$) SNP. For simplicity we assume that X_{ij} denotes whether the rare allele A_j is present ($X_{ij} = 1$) or absent ($X_{ij} = 0$) in person i and let $\mathbf{X}_i = (X_{i1}, \dots, X_{iJ})'$. It is straightforward to consider the case where X_{ij} is the number (0, 1 or 2) of the rare allele for SNP j , but since A_j is rare there will be no or very few individuals with genotype A_jA_j in a study of typical size. Tests below can also be readily modified to accommodate vectors X_{ij} that code genotypes A_jA_j , A_jB_j or B_jB_j in other ways. We assume for now that there is no adjustment for covariates; we address this in Section 7.

Most proposed methods (e.g. see Basu and Pan (2011)) for testing a null hypothesis of no association between Y and \mathbf{X} are based on statistics S_j ($j = 1, \dots, J$), which individually measure association between Y and given SNP j . Without loss of generality we assume that S_j is such that $E[S_j] = 0$ and $\text{Var}(S_j) = \sigma_{0j}^2$ under the null hypothesis; under alternatives, we denote $E[S_j]$ and $\text{Var}(S_j)$ by μ_j and σ_j^2 . There are many options for S_j ; for example $S_j = \sum_{i=1}^n (Y_i - \bar{Y})X_{ij}$ is natural choice if the Y_i are approximately normally distributed and is also used with binary traits. The

approaches referred to in Section 1 can be expressed in terms of statistics of the form

$$S_j = \sum_{i=1}^n \alpha_i X_{ij}, \quad j = 1, \dots, J \quad (2.1)$$

where α_i is a function of either Y_i or its rank, with $\sum_{i=1}^n \alpha_i = 0$. Such statistics arise naturally from regression models relating Y and \mathbf{X} , as we discuss later.

Our interest is in testing the null hypothesis

$$H_0 : Y \text{ and } \mathbf{X} \text{ are independent.} \quad (2.2)$$

We first review the permutation distribution of $\mathbf{S} = (S_1, \dots, S_J)'$. Although exact or asymptotic model-based distributions can be obtained in many cases, the permutation distribution is often used to compute p-values; this is the distribution that arises from randomly permuting Y_1, \dots, Y_n and assigning them to the \mathbf{X}_i . Under H_0 , the permutation mean of \mathbf{S} is $\mathbf{0}$ and the covariance matrix Σ_S has entries (e.g. Kalbfleisch and Prentice (2002), Sec. 7.2)

$$\begin{aligned} \Sigma_P(j, \ell) = \text{cov}_P(S_j, S_\ell) &= \left(\frac{\sum_{i=1}^n \alpha_i^2}{n-1} \right) \left(\sum_{i=1}^n (X_{ij} - \bar{X}_j) (X_{i\ell} - \bar{X}_\ell) \right) = \\ &= \left(\frac{\sum_{i=1}^n \alpha_i^2}{n-1} \right) \left(m_{jl} - \frac{m_j m_\ell}{n} \right), \end{aligned} \quad (2.3)$$

where $m_j = \sum_{i=1}^n X_{ij}$ and $m_{jl} = \sum_{i=1}^n X_{ij} X_{il}$, for $j = 1, \dots, J$ and $\ell = 1, \dots, J$. This also applies when Y is a discrete variable (see supplementary materials), when the genotypes X_{ij} are correlated within individuals (e.g. due to linkage disequilibrium, LD) and when sampling of individuals is Y -dependent.

Many authors have considered linear test statistics for H_0 of the form

$$W_L = \sum_{j=1}^J w_j S_j = \mathbf{w}' \mathbf{S}, \quad (2.4)$$

where the weights w_j are specified positive values and $\mathbf{w} = (w_1, \dots, w_J)'$. Basu and Pan (2011) provide a review, but we note two cases: Morgenthaler and Thilly (2007) considered the “cohort allelic sums test” (CAST) where each $w_j = 1$, and Madsen and Browning (2009) based w_j on the population MAF p_j , with larger weights for SNPs with smaller p_j . The rationale for the latter weights is that deleterious SNPs would be subject to “purifying selection” and so be rarer in the population than neutral SNPs, but evidence for this so far seems slight. Price et al. (2010) also considered “threshold” versions in which $w_j > 0$ only if the relative frequency of SNP j is below a specified threshold (e.g. 1% or 5%). Such statistics can have good power against alternatives where $E[S_j] = \mu_j \geq 0$, with $\mu_j > 0$ for some subset of $\{j = 1, \dots, J\}$. However, their power may be poor for alternatives where both positive and negative values of μ_j are possible and, in addition, when only a small proportion of the J SNPs have $\mu_j > 0$ (Neale et al. (2011), Basu and Pan (2011)). This is studied here in Sections 4 and 5.

To facilitate further discussion, we assume without loss of generality that Y is defined so that a SNP with $\mu_j > 0$ is termed deleterious (harmful) and one with $\mu_j < 0$ is termed protective (beneficial). A SNP with $\mu_j = 0$ is termed neutral and deleterious or protective SNPs are termed causal. In the absence of prior knowledge concerning the effects of SNPs, a preferable family of statistics might be of quadratic form $W_Q = \mathbf{S}' \mathbf{A} \mathbf{S}$, where \mathbf{A} is a positive definite (or semi-definite) symmetric matrix.

A common choice is $A = \Sigma_S^{-1}$ and this is effectively a Hotelling statistic,

$$W_H = \mathbf{S}' \Sigma_S^{-1} \mathbf{S}, \quad (2.5)$$

where Σ_S is the covariance matrix of \mathbf{S} under H_0 . This statistic arises from models for Y given \mathbf{X} in many settings, and we consider it in theoretical and numerical studies below.

Basu and Pan (2011), Neale et al. (2011), Wu et al. (2011), Lee et al. (2012) and others have considered other quadratic statistics,

$$W_Q = \mathbf{S}' A \mathbf{S} \quad (2.6)$$

For example, the 'SSU' statistic (Pan (2009)) and 'C-alpha' statistic of Neale et al. (2011) are based on $A = I$, the $J \times J$ identity matrix and the SKAT statistic of Wu et al. (2011) uses $A = \text{diag}(w_1, \dots, w_J)$, where the w_j are weights. Linear statistics W_L in (2.4) can also be expressed in this form, since W_L^2 is given by (2.6) with $A = \mathbf{w}'\mathbf{w}$.

Statistics of the form (2.6) can be obtained from random effect regression models in which Y is related to \mathbf{X} through a linear function $\beta' \mathbf{X}$ and the $J \times 1$ regression coefficient β is a random vector with mean $\mathbf{0}$ and covariance matrix τA . The hypothesis $\tau = 0$ then corresponds to (2.2) and a score statistic for testing it is (Goeman et al. (2006), Basu and Pan (2011))

$$W'_Q = \frac{1}{2} \mathbf{S}' A \mathbf{S} - \frac{1}{2} \text{trace}(A \Sigma_S).$$

Using W'_Q is equivalent to using (2.6). A number of authors have claimed that as an "omnibus" test statistic, the Hotelling statistic (2.5) lacks power in many settings. Conversely, the fact that (2.6) can be obtained as a score test for single parameter (τ)

has suggested to many that it will have more power than (2.6). It is often overlooked, however, that the choice $A = \Sigma_S^{-1}$ in (2.6) produces (2.5) and so it can be obtained as a “single parameter” test. The performance of a test statistic (2.6) depends on A and on the mean $\boldsymbol{\mu}$ and covariance matrix Σ_S of S under alternative hypotheses and we explore this in what follows.

It is instructive to consider the case where \boldsymbol{S} is normally distributed. The vectors \boldsymbol{S} considered here and by others are all at least asymptotically normal, and analytical derivations of power and discussions of optimality (e.g. Lin and Tang (2011); Lee et al. (2012)) rely on this. The normal case is well known in multivariate analysis, in connection with tests for a multivariate mean $\boldsymbol{\mu} = E(\boldsymbol{S})$, where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_J)'$; see for example Mardia et al. (1979), Ch. 5. For tests of $H_0: \boldsymbol{\mu} = \mathbf{0}$, the power of statistics such as (2.4)-(2.6) against an alternative hypothesis H_1 for which $\boldsymbol{\mu} \neq \mathbf{0}$ depends on $\boldsymbol{\mu}$ and on the distribution of \boldsymbol{S} under H_1 . In particular, suppose that under H_1 the distribution of \boldsymbol{S} is multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ . For simplicity we assume Σ is known; this is the case for some statistics and asymptotically it is generally all right to assume it. The following distributional results then hold: let $\lambda_1, \dots, \lambda_J$ be the eigenvalues of $\Sigma^{1/2} A \Sigma^{1/2}$ and let P be the $J \times J$ orthogonal matrix whose columns are the corresponding eigenvectors. Then

- (i) W_Q is distributed as a linear combination of independent non-central χ^2 random variables,

$$W_Q \sim \sum_{j=1}^J \lambda_j \chi_{1, nc_j}^2 \quad (2.7)$$

where $nc_j = (P' \Sigma^{-1/2} \boldsymbol{\mu})_j^2$ and $\chi_{k,r}^2$ denotes a non-central χ^2 random variable with k degrees of freedom and non-centrality parameter r .

- (ii) Under the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$, W_Q is a linear combination of independent χ_1^2 random variables; each $nc_j = 0$ in (2.7).

(iii) If $A = \Sigma^{-1}$ then $W_Q \sim \chi_{J,nc}^2$, where $nc = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$. In this case $W_Q = W_H$ of (2.5).

(iv) $Z_L^2 = (\mathbf{w}'\mathbf{S})^2/(\mathbf{w}'\Sigma\mathbf{w})$ is distributed as $\chi_{1,nc}^2$ with $nc = (\mathbf{w}'\boldsymbol{\mu})^2/(\mathbf{w}'\Sigma\mathbf{w})$.

For such distributional results, see Rao (1973), Sec. 3b.4. We note that software exists for the computation of probabilities associated with linear combinations of central or non-central χ_1^2 random variables. In particular, we used the *CompQuadForm* package in R (Duchesne and de Micheaux (2010)).

The results above allow the power against an alternative hypothesis where $\mathbf{S} \sim N(\boldsymbol{\mu}, \Sigma)$ to be calculated for any linear test statistic (2.4) or quadratic test statistic (2.6). Critical values for a test of $H_0: \boldsymbol{\mu} = \mathbf{0}$ are obtained according to (ii). We note in particular that

- (a) For a (two - sided) size α test using the linear statistic (2.4) or equivalently Z_L^2 in (iv) above, the α critical value for Z_L^2 is $\chi_1^2(1 - \alpha)$, the $1 - \alpha$ quantile for the χ_1^2 distribution. The power against an alternative $(\boldsymbol{\mu}, \Sigma)$ is then

$$P(\chi_{1,nc_L}^2 > \chi_1^2(1 - \alpha))$$

where $nc_L = (\mathbf{w}'\boldsymbol{\mu})^2/(\mathbf{w}'\Sigma\mathbf{w})$

- (b) For a size α test using the Hotelling statistic (2.5), the α critical value for W_H is $\chi_J^2(1 - \alpha)$. The power against an alternative $(\boldsymbol{\mu}, \Sigma)$ is

$$P(\chi_{J,nc_H}^2 > \chi_J^2(1 - \alpha))$$

where $nc_H = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$.

A number of authors have claimed to obtain “optimal” tests. This is theoretic-

cally possible if we specify a suitable family of test statistics but for this to be of practical use we must have strong prior knowledge about the alternative hypothesis. For example, among the family of linear statistics (2.4), maximal power is obtained when $\mathbf{w} = \Sigma^{-1}\boldsymbol{\mu}$; when the S_j are independent so that $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_J^2)$, this gives $w_j = \mu_j/\sigma_j^2$ ($j = 1, \dots, J$). This linear statistic is in fact optimal among all tests of fixed size based on \mathbf{S} . Quadratic statistics (2.5) or (2.6) for which \mathbf{A} has rank 2 or more can never be optimal against a specific alternative $(\boldsymbol{\mu}, \Sigma)$. However, they can maintain high power over wide ranges of alternatives, whereas a linear statistic's power can be poor except near a specific alternative. Goeman et al. (2006) and others have discussed optimality of score statistics (2.6) coming from random effects models, but these results are based on averaging over a family of alternatives, which may or may not be plausible in a given setting.

The Hotelling statistic (2.5) is a reasonable choice when alternatives with both deleterious ($\mu_j > 0$) and protective ($\mu_j < 0$) SNPs are plausible and also performs well more generally, as we show later. It can be shown that for a given linear statistic $W_L = \mathbf{w}'\mathbf{S}$ we can decompose W_H as

$$W_H = Z_L^2(w) + Q(w) \quad (2.8)$$

where $Q(w)$ and $Z_L(w)$ are independent and, under an alternative $(\boldsymbol{\mu}, \Sigma)$, $Q(w)$ is non-central χ_{J-1, nc_Q}^2 with

$$nc_Q = nc_H - nc_L = \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} - (\mathbf{w}'\boldsymbol{\mu})^2/(\mathbf{w}'\Sigma\mathbf{w}). \quad (2.9)$$

The linear statistic will have low power when $nc_Q \geq 0$ is large, while it is optimal when $nc_Q = 0$.

The test statistic $W_C = \mathbf{S}'\mathbf{S}$, given by (2.6) with $\mathbf{A} = \mathbf{I}$, has been found powerful

in a quite wide range of settings (e.g. Neale et al. (2011), Basu and Pan (2011)). For the most part, the settings investigated were ones where the regression coefficients β_j in a model for Y given \mathbf{X} were unrelated to the frequency (MAF) of the rare variant. In cases where causal SNPs (and larger $|\beta_j|$) are more likely to be found among rarer variants, the situation may, however, be reversed. To illustrate this, let $p_j = P(X_{ij} = 1)$ in population and suppose the individuals represent a random sample (Y_i, \mathbf{X}_i) , $i = 1, \dots, n$. The covariance matrix Σ_S in (2.3) is approximately a multiple of the diagonal matrix $\text{diag}\{p_1(1 - p_1), \dots, p_J(1 - p_J)\}$ when the SNPs are mutually independent. Then, (2.5) is approximately $\sum_{j=1}^J \frac{S_j^2}{p_j(1-p_j)}$ so that S_j is weighted inversely according to the MAF of SNP j . It is thus possible that W_H may be more powerful than $W_C = \sum_{j=1}^J S_j^2$. We investigate this further in Sections 4 and 5.

Finally, some authors (e.g. Han and Pan (2010), Hoffmann et al. (2010), Lin and Tang (2011)) have proposed two-stage or other adaptive approaches in which the weight vector \mathbf{w} in (2.4) is chosen after preliminary examination of the direction of S_j or an estimate of its effect. However, such approach cannot on its own (i.e. without the use of background information from other sources) improve globally on quadratic statistics. In fact, if we choose the \mathbf{w} that maximizes the standardized linear test statistic (2.4), then we end up with the quadratic statistic (2.5). In particular (see e.g. Mardia et al. (1979) p. 127 or Li and Lagakos (2006), Sec. 3),

$$\sup_{\mathbf{w}} \left\{ \frac{W_L^2}{\text{Var}(W_L)} \right\} = \sup_{\mathbf{w}} \left\{ \frac{(\mathbf{w}' \mathbf{S})^2}{\mathbf{w}' \Sigma_S \mathbf{w}} \right\} = \mathbf{S}' \Sigma_S^{-1} \mathbf{S} = W_H,$$

where the maximizing vector is $\mathbf{w} = \Sigma_S^{-1} \mathbf{S}$. This helps explain why Basu and Pan (2011) found that adaptive procedures did not perform as well as one might hope.

We remark that Lin and Tang (2011) have proposed a test statistic T_{max} based on the maximum of a specified set of K linear statistics, each with different weights,

$T_k^2 = (\mathbf{w}_k' \mathbf{S})^2 / (\mathbf{w}_k' \Sigma_S \mathbf{w}_k)$. We do not consider such statistics here, but it is clear that their performance will depend on the choice of ‘appropriate’ weight vectors \mathbf{w}_k . In the case where there is little prior information and the \mathbf{w}_k are selected to cover a range of alternatives with deleterious and protective SNPs, it seems likely that $\max(T_k^2)$ would be similar to W_H . A similar suggestion involving quadratic statistics is made by Lee et al. (2012).

In practice there is often very limited prior information about the nature of $\boldsymbol{\mu}$, especially concerning which SNPs might be causal, so one cannot be confident that a linear test statistic (2.4) will be effective, nor which quadratic statistics might be best. In Sections 4 and 5, we investigate situations in which a specific statistic will be more or less powerful.

3 Dichotomous Traits

Many previous investigations have focused on binary responses and case-control sampling. Suppose $Y_i = 1$ and $Y_i = 0$, respectively, indicate whether an individual has a certain condition (“case”) or not (“control”). A widely used statistic for assessing association between binary Y_i and X_{ij} is (e.g. Basu and Pan (2011), Neale et al. (2011))

$$S_j = \sum_{i=1}^n (Y_i - \bar{Y}) X_{ij} = T_j - n_1 m_j / n, \quad (3.1)$$

where $T_j = \sum_{i=1}^n X_{ij} Y_i$, $m_j = \sum_{i=1}^n X_{ij}$ and $n_1 = \sum_{i=1}^n Y_i$. When X_{ij} denotes the presence ($X_{ij} = 1$) or absence ($X_{ij} = 0$) of the rare variant for SNP j , T_j is the number of cases with the rare variant and m_j is the total number of individuals with the rare variant; n_1 is the number of cases. The statistic $\mathbf{S} = (S_1, \dots, S_J)'$ has expectation

zero under H_0 for either random or case-control sampling; when cases are rare in the population, the latter sampling design is typically used.

Under either type of sampling, the conditional distribution of S_j given m_j , n and n_1 , and under H_0 , is hypergeometric,

$$\Pr(T_j = t_j) = \binom{n_1}{t_j} \binom{n_0}{m_j - t_j} / \binom{n}{m_j}, \quad (3.2)$$

where $n_0 = n - n_1$. This is also the permutation distribution for T_j when the n_1 cases and n_0 controls in the sample are randomly assigned to individuals $1, \dots, n$, and it is the basis of Fisher's exact test applied to the 2×2 table defined by $X_{ij} = 0$ or 1 and $Y_i = 0$ or 1 . In our case we want to consider all J SNPs, and we find from (2.3) that the covariance matrix for \mathbf{S} has entries

$$\Sigma_P(j, \ell) = \frac{n_0 n_1}{n(n-1)} \left(m_{j\ell} - \frac{m_j m_\ell}{n} \right), \quad (3.3)$$

for $j = 1, \dots, J$; $\ell = 1, \dots, J$, where $m_{j\ell} = \sum_{i=1}^n X_{ij} X_{i\ell}$ is the number of individuals with the rare variant for both SNP j and SNP ℓ . The diagonal entries in (3.3) agree with variances obtained from (3.2), but the T_j may not be (conditionally) independent for $j = 1, \dots, J$.

As in Section 2, choice of a statistic for testing H_0 should be guided by which alternatives for μ_j ($j = 1, \dots, J$) are of interest. Linear statistics of the form (2.4) have been considered in case-control settings by several authors cited above. Recently, Neale et al. consider the quadratic test statistic

$$W'_C = \sum_{j=1}^J \left(S_j^2 - \frac{n_0 n_1 m_j}{n^2} \right). \quad (3.4)$$

This is, for large n and small m_j/n , approximately equal to

$$\begin{aligned} W_C'' &= \sum_{j=1}^J (S_j^2 - \text{Var}_P(S_j)) \\ &= \mathbf{S}'\mathbf{S} - \text{trace}(\Sigma_P). \end{aligned} \tag{3.5}$$

This test arises from a random effects model as discussed in Section 2 and is also a "C-alpha"(score) test for extra-binomial variation in the T_j . Basu and Pan (2011) noted the essential equivalence of W_C'' to W_Q' , and to $W_C = \mathbf{S}'\mathbf{S}$.

The distributions of test statistics such as (2.4), (2.5) or (3.4) under null (H_0) and alternative (H_1) hypotheses can be approximated in sufficiently large samples by normal, chi-squared, or linear combination of chi-square distributions. This follows directly from the fact that $n^{-1/2}\mathbf{S}$ is asymptotically normal as n goes to infinity with n_1/n and n_0/n fixed, and an application of results in Section 2. Although the approximations may be inadequate in some situations, examination of power under (approximate) normality of \mathbf{S} provides considerable insights for both quantitative and binary traits. Thus we next assume normality for \mathbf{S} in Section 4 and provide analytical power comparisons for quantitative traits. We then return to case-control settings in Section 5 and empirically evaluate the performance of test statistics via simulation.

4 Power Calculations Under Normality

4.1 Theoretical calculations

We consider a setting where X_{ij} denotes whether individuals i has ($X_{ij} = 1$) or does not have ($X_{ij} = 0$) the rare variant of SNP j , and assume a normal linear model.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_J X_{iJ} + e_i \quad \text{for } i = 1 \dots n \quad (4.1)$$

with $e_i \sim N(0, \sigma^2)$ and the X_{ij} mutually independent Bernoulli variables with $P(X_{ij} = 1) = p_j$ ($j = 1, \dots, J$). Explained variation of the J SNPs under model (4.1) is

$$EV = \frac{\text{Var}(E(Y|\mathbf{X}))}{\text{Var}(Y)} = \frac{\sum_{j=1}^J p_j(1-p_j)\beta_j^2}{\sum_{j=1}^J p_j(1-p_j)\beta_j^2 + \sigma^2} \approx \sum_{j=1}^J \frac{p_j(1-p_j)\beta_j^2}{\sigma^2} = \sum_{j=1}^J EV_j$$

when EV is small. We will refer to $EV_j = p_j(1-p_j)\beta_j^2/\sigma^2$ as the “explained variation” due to SNP j . The score statistic $\mathbf{S} = (S_1, \dots, S_J)'$ with

$$S_j = \sum_{i=1}^n (Y_i - \bar{Y})X_{ij} = \sum_{i=1}^n (X_{ij} - \bar{X}_j)Y_i, \quad (4.2)$$

arises from maximum likelihood theory for testing $H_0 : \boldsymbol{\beta} = (\beta_1, \dots, \beta_J)' = \mathbf{0}$. Due to normality of the Y_i , the distribution of S_j given the \mathbf{X}_i ($i = 1, \dots, n$) is

$$S_j \sim N(A_j\beta_j, A_j\sigma^2) \quad (4.3)$$

where $A_j = m_j(1 - m_j/n)$. For simplicity we consider the case where m_j and m_{jl} equal their expected values, $np_j(1 - p_j)$ and np_jp_l ($j \neq l$), so that

$$\mathbf{S} \sim N(\boldsymbol{\mu}, \Sigma_S). \quad (4.4)$$

where $\boldsymbol{\mu} = (np_1(1-p_1)\beta_1, \dots, np_J(1-p_J)\beta_J)'$ and $\Sigma_S = \text{diag}\{np_1(1-p_1)\sigma^2, \dots, np_J(1-p_J)\sigma^2\}$.

We consider two linear statistics (2.4) : W_{LS} with weights w_j set to 1 and W_{LW} with weights w_j set to $1/\sqrt{(p_j(1-p_j))}$. We also consider two quadratic statistics (2.6) : W_C with matrix $A = I$ (C-alpha) and W_H with matrix $A = \Sigma_S^{-1}$ (Hotelling). Under model (4.1) we have from the results in Section 2 that

- i) W_{LS} is distributed as $W_{LS} \sim N(\sum_{j=1}^J np_j(1-p_j)\beta_j, \sum_{j=1}^J np_j(1-p_j)\sigma^2)$
- ii) W_{LW} is distributed as $W_{LW} \sim N(\sum_{j=1}^J n\sqrt{p_j(1-p_j)}\beta_j, nJ\sigma^2)$
- iii) W_C is distributed as $W_C \sim \sum_{j=1}^J \lambda_j \chi_{1,nc_j}^2$, where $\lambda_j = np_j(1-p_j)\sigma^2$ and $nc_j = np_j(1-p_j)\beta_j^2/\sigma^2 = nEV_j$.
- iv) W_H is distributed as $W_H \sim \chi_{J,nc}^2$, where $nc = \sum_{j=1}^J np_j(1-p_j)\beta_j^2/\sigma^2$, which is approximately equal to nEV .

Thus the power of W_H depends (approximately) just on the total explained variation and sample size and is not sensitive to the specific effects of the causal variants. On the other hand the power of the C-alpha statistic depends on the explained variation of the SNPs and also corresponding “weights” λ_j . The effect of the λ_j is to give larger weight to more common SNPs (SNPs with larger MAF). The power of a linear statistic W_L (2.4) is a function of the non-centrality parameter

$$nc_L(\mathbf{w}) = \frac{n}{\sigma^2} \frac{(\sum_{j=1}^J w_j p_j (1-p_j) \beta_j)^2}{\sum_{j=1}^J w_j^2 p_j (1-p_j)}. \quad (4.5)$$

To get weights near the optimal $w_j = \beta_j/\sigma^2$ requires significant prior information, and the effect of weights inversely proportional to p_j is unclear. We provide numerical results on power of W_{LS} , W_{LW} , W_C and W_H under various conditions in the next section.

4.2 Numerical Comparisons

In order to consider a broad range of scenarios, we randomly generated 1000 different models. Each model has randomly selected values for J , J_C , J_D , β_j and p_j as follows: J - number of SNPs was randomly selected from the set $\{10, 20, 30, 40, 50\}$; $p_C = J_C/J$ - proportion of causal SNPs was generated from $U(0.1, 1)$; $p_D = J_D/J_C$ - proportion of deleterious SNPs was generated from $U(0.75, 1)$, β_j - genetic effect of j^{th} SNP was generated differently depending on the scenarios described later and p_j - probability of the rare variant for the j^{th} SNP was generated from $U(0.005, 0.02)$. Sample size was fixed to be $n = 1000$ and the level of the test was set to be $\alpha = 0.0001$ to reflect the fact that testing would typically be conducted for multiple genetic regions. Since the power of W_{LS} , W_{LW} , W_C and W_H is a function of values β_j/σ and p_j , without loss of generality we set $\sigma^2 = 1$. We consider two different situations concerning p_j and genetic effect β_j . The first scenario assumes no relationship between p_j and β_j for each SNP and takes $|\beta_j| \sim U(0.45, 0.5)$ for each causal SNP. Explained variation EV_j of a single causal SNP ranges between 0.1% and 0.49% under this scenario with SNPs with smaller MAF having smaller explained variation. The second scenario assumes that explained variation due to a single causal SNP is independent of MAF and therefore SNPs with smaller MAFs have larger genetic effects β_j . Explained variation EV_j of a single causal SNP is generated from $U(0.001, 0.0025)$ in this case, and the genetic effect β_j is then determined from $EV_j = \frac{p_j(1-p_j)\beta_j^2}{\sigma^2}$.

Figure 1 shows results based on the 1000 randomly generated models under the first scenario. It compares the analytically calculated power of linear statistics with and without weighting, W_{LS} and W_{LW} . It also compares power of the quadratic Hotelling statistic W_H and C-alpha statistic W_C . In view of the wide variations in model parameters, the powers of the tests vary widely across the 1000 models, but that the power of the two linear statistics is similar for each model, and the power of the

two quadratic statistics is similar for each model. Moreover, neither statistic within each class dominates across a majority of the models. Figure 2 has results based on the 1000 randomly generated models under the second scenario. We now see that for the linear statistics the picture is similar to that in Figure 1, but the Hotelling statistic performs better than the C-alpha statistic across almost all models, as our earlier comments suggest.

Figures 3 and 4 compare linear and quadratic statistics. Figure 3 compares power of the linear statistic W_{LS} and the Hotelling statistic W_H for 1000 models generated under the first scenario, and Figure 4 compares power of W_{LS} and W_H for 1000 models generated under the second scenario. In this case we consider three sample sizes: $n = 500, 1000$ and 2000 . Figure 3 indicates that with sample size $n = 500$ there is low power for both statistics in a large fraction of the 1000 scenarios. The linear statistic more often achieves a moderately high power, in essence because only models with fairly high proportions of causal SNPs with similar (same direction) effects produce much power. As n increases, however, the quadratic statistic displays good power across many models and by $n = 2000$ dominates the linear statistic for most of the models. In Figure 4 the linear statistic dominates when $n = 500$, but power exceeds 0.5 in only small proportion of models. When $n = 1000$ the linear and quadratic statistics are best about equally often, but the linear statistic achieves high power more often. When $n = 2000$, however, the quadratic statistic dominates in the vast majority of the scenarios.

We also investigated the relationships between model parameters and the power of the linear (W_{LS}) and Hotelling (W_H) statistics. We present results for settings with $J = 30$ and $n = 1000$ in Figure 5. Results are based on the 10,000 randomly generated models under the first scenario. Figure 6 has results based on the second scenario, with explained variation of each SNP independent of p_j . To show the relationship

between power of the tests and the number of causal SNPs we grouped models by the number of causal SNPs; to show the relationship between power and the proportion of deleterious SNPs, we sub-grouped models by the number of deleterious SNPs. The X axis in Figures 5 and 6 indicates both the number J_C of causal SNPs and the number J_D of deleterious SNPs in each model. Values $J_C = 3, 4, \dots, 30$ give the main scale in each figure, with J_D ranging from $0.75J_C$ up to J_C between successive values for J_C .

Figures 5 and 6 suggest a number of important conclusions:

- i) Performance of the linear statistic varies widely across scenarios. For W_{LS} to perform well it is necessary not only that the effects of causal SNPs are (almost) all in the same direction (deleterious or protective), but also that the proportion of causal to neutral SNPs is not too low.
- ii) The quadratic statistic W_H performs well across a range of scenarios with varying proportions of deleterious, protective and neutral SNPs. It can outperform linear statistics even in cases when causal SNP effects are all in the same direction, if the ratio of causal to neutral SNPs is not high.
- iii) The powers of both W_{LS} and W_H strongly depend on the percentage of causal SNPs in the group of SNPs. For settings here with $n = 1000$ and realistic levels of explained variation for causal SNPs, powers of 0.5 or more require that a majority of the SNPs be causal. The results suggest that considerable attention should be given to background information that can guide selection of SNPs for pooled testing, and that rather large sample sizes may be needed to provide adequate power.

Finally, we generated J_D/J_C from $U(0.75, 1)$ here to reflect the common assumption that rare causal SNPs are more likely to be deleterious than protective. However,

we also simulated models where J_D/J_C was $U(0.5, 0.75)$. In that case the linear statistics performed poorly and, as one would expect, were dominated by the quadratic statistics.

5 Simulation Studies for Case-Control Settings

Here, we provide detailed numerical results for case-control studies when a normal approximation for \mathbf{S} might not be adequate. As in previous sections, we examine the performance of W_{LS} , W_{LW} , W_C and W_H . We first considered a normal approximation for the linear statistics and linear combination of chi-squares (see (2.7)) for quadratic statistics for obtaining p-values or Type I errors, and investigated their adequacy by simulation. We then conducted simulations to assess the statistics power under different scenarios.

We assume that the distribution of Y_i given \mathbf{X}_i in the population is Bernoulli with

$$p(\mathbf{X}_i) = P(Y_i = 1 | \mathbf{X}_i) = \frac{\exp(\beta_0 + \sum \beta_j X_{ij})}{1 + \exp(\beta_0 + \sum \beta_j X_{ij})} \quad (5.1)$$

and that the X_{ij} in the population are mutually independent Bernoulli variables with $P(X_{ij} = 1) = p_j$ for $j = 1, \dots, J$. Similar to Section 4, we consider for power assessments a broad range of scenarios by randomly generating 500 different models. Each model has randomly selected values J , J_C , J_D , β_j and p_j as follows: J was randomly selected from the set $\{10, 20, 30, 40, 50\}$; $p_C = J_C/J$ was generated from $U(0.1, 1)$; $p_D = J_D/J_C$ was generated from $U(0.75, 1)$, β_j was generated differently depending on the scenarios described later and p_j was generated from $U(0.005, 0.02)$. Sample size for both cases and controls was fixed to be $n_1 = n_0 = 500$, the level of the test was set to be $\alpha = 0.0001$ and without loss of generality we took $\beta_0 = -2.1922$

(giving $P(Y_i = 1 | \mathbf{X}_i = \mathbf{0}) = 0.1$). We consider two different situations concerning p_j and β_j , which is a log odds ratio, $\log(\text{OR})$. The first scenario assumes no relationship between p_j and β_j for each SNP and takes $e^{|\beta_j|} \sim U(1.5, 3)$ for each causal SNP. The second scenario assumes that the odds ratio for a causal SNP is inversely proportional to $\sqrt{p_j(1 - p_j)}$; we set $e^{|\beta_j|} = C/\sqrt{p_j(1 - p_j)}$, where $C = 4\sqrt{0.005(1 - 0.005)}$. Under these scenarios the odds ratio for deleterious variants ranges from 2 for $p_j = 0.02$ to 4 for $p_j = 0.005$ and from 0.5 down to 0.25 for protective variants.

Due to the sparsity of rare variants and the dichotomous trait, normal and chi-square approximations for statistics may produce inflated or conservative Type I errors as was observed in previous studies (Lin and Tang (2011), Basu and Pan (2011)). To investigate the suitability of the approximations for nominal Type I error $\alpha = 0.05, 0.01, 0.001$ and 0.0001 , we considered the model with $p_j = 0.01$ for all J SNPs under null hypothesis ($\beta = 0$). Results are displayed in Table 1 for linear statistic W_{LS} and Hotelling statistic W_H (Table 1). Empirical Type I errors in the table are based on 10^6 simulation replicates, and are the proportions of the 10^6 samples for which the corresponding test statistic exceeded the α critical values given by a normal or chi-square approximation. The 10^6 simulation replicates could be realized efficiently because the S_j are mutually independent and are functions of the T_j and m_j only; under H_0 , m_j is generated from $\text{Binomial}(1000, 0.01)$ and T_j from (3.2). Table 1 shows that the normal approximation for W_L is accurate, but quadratic tests have conservative Type I errors with chi-square approximations. Therefore, for each randomly chosen scenario for assessing power, we first generated 10^6 models under the null hypothesis ($\beta = 0$), to obtain empirical critical values for Type I error $\alpha = 0.0001$.

The case-control samples under H_1 were simulated using for computational convenience the assumption that the X_{ij} are mutually independent, given $Y_i = 0$ or $Y_i = 1$.

Under this assumption, $\mathbf{X}_i = (X_{i1}, \dots, X_{iJ})'$, for the i^{th} individual with case status is generated from $X_{ij} = \text{Binomial}(1, p(X_{ij} = 1|Y_i = 1))$. Similarly, $\mathbf{X}_i = (X_{i1}, \dots, X_{iJ})'$ for i^{th} individual with control status is generated from $X_{ij} = \text{Binomial}(1, p(X_{ij} = 1|Y_i = 0))$. Conditional probabilities of $X_{ij} = 1$ given $Y_i = 0$ or $Y_i = 1$ were calculated using Monte-Carlo sampling. This is an approximation, but empirical powers were found to be very close to those obtained under much more computationally intensive simulations that are based on the exact model. For each combination of parameters, simulations to assess power consisted of 1,000 replications, with power estimated by the proportion of samples for which test statistic exceeded its critical value.

Figure 7 shows results based on the 500 randomly generated models under the first scenario. It compares the empirically calculated power for linear statistics with and without weighting, W_{LS} and W_{LW} . It also compares power of the Hotelling statistic W_H and C-alpha statistic W_C . We see that the C-alpha performs better than the Hotelling statistic, while the linear statistic with no weights has an advantage over the statistic with weights. Under case-control scenarios, deleterious SNPs have larger values of $m_j = \sum_{i=1}^n X_{ij}$ than neutral SNPs with the same p_j due to enrichment from the cases. This explains why W_{LS} and W_C have an advantage over W_{LW} and W_H , respectively. Similar results were obtained by Basu and Pan (2011).

Figure 8 shows results based on the 500 randomly generated models under the second scenario. We again see a difference in power between linear statistics W_{LS} and W_{LW} . However, the systematic power difference between the quadratic statistics is absent in Figure 8. This supplements the results of Basu and Pan (2011), who did not consider cases where the genetic effect is inversely proportional to MAF, and indicates that the relative performance of W_C and W_H depends on the relationship between SNP effects and MAF.

We also compare power of the linear statistic W_{LS} and quadratic statistic W_C for

the 500 models under the first scenario in Figure 9, and Figure 10 compares power of those statistics under the second scenario. Similar conclusions to those for the quantitative trait study can be made. The variation in power between the two tests is quite large here but in Figure 10 especially, the linear statistics achieves high power more often. We also investigated the relationships between model parameters and the power of the linear (W_{LS}) and quadratic (W_C) statistics. We present results for settings with $J = 30$ and $n_1 = n_0 = 500$ in Figure 11 under assumption of independent of β_j and p_j and in Figure 12 under the odds ratio for a causal SNP is inversely proportional to $\sqrt{p_j(1 - p_j)}$. Similarly to numerical comparisons under normality, we observed that linear statistics have lower power than quadratic statistics even when all or almost all causal SNPs have effect in the same direction. One difference from results in Section 4 is that linear statistics has smaller drop of power when large proportion of causal SNPs have effect in the different direction. This is because the association statistics S_j in the case-control setting is scaled MAF difference between cases and controls. Under the logistic model, this difference is not symmetric for deleterious and protective SNPs even when they have same effect but different sign.

6 Applications: GAW17 Data

The numerical studies in Sections 4 and 5 assume that SNPs were mutually independent. To consider settings where this might not be so, we examined real human sequence data (1000 Genomes Project Consortium (2010)) that were used to generate quantitative trait data in Genetic Analysis Workshop 17 (GAW 17); see Almsay et al. (2011).

We analyzed quantitative trait Q2 which was influenced by 72 SNPs (70 with MAF in the range (0.16%, 17%)) in 13 genes (see Table 2 of Almsay et al. (2011)) but not

by other covariates . We used data from the $n = 321$ unrelated Asian subjects (Han Chinese, Denver Chinese and Japanese). We calculated permutation-based p-values for the four tests, W_{LS} , W_{LW} , W_C and W_H discussed in previous sections, and we estimated power by analyzing all 200 replicates provided by GAW17 at the $\alpha = 0.05$ level (Table 2). The choice $\alpha = 0.05$ was based on the overall low power for detecting effects due to small sample size, small genetic effect, extremely small MAF and the low proportion of causal variants in a gene.

Results in Table 2 are consistent with our previous conclusions: (i) linear tests with and without weighting based on MAF vary in relative power; (ii) quadratic statistics W_C and W_H also have variable relative power; (iii) relative performance of the linear and quadratic tests is highly variable. As expected, with a large number of causal SNPs linear statistics can outperform quadratic (e.g. genes SIRT1, SREBF1), but when the proportion of causal SNPs is low quadratic statistics outperform linear (e.g. genes BCHE, PDGFD, RARB).

7 Regression Models and Additional Covariates

In some settings a test of no association may be based on a regression model (e.g. Morris and Zeggini (2010)). In fact, Lin and Tang (2011) and Wu et al. (2011) have stressed that adjustment for covariates and population stratification will be important in many contexts involving rare variants. We now discuss how our framework is extended to deal with the inclusion of covariates; for illustration, we consider the case of a binary trait. Suppose that in addition to the genotype vector \mathbf{X}_i there is a vector \mathbf{v}_i of covariates that may be related to a binary trait Y_i . Then a logistic regression model

$$\Pr(Y_i = 1 | \mathbf{X}_i, \mathbf{v}_i) = \frac{\exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + \boldsymbol{\gamma}'\mathbf{v}_i)}{1 + \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + \boldsymbol{\gamma}'\mathbf{v}_i)} = \mu_i \quad (7.1)$$

might be considered, and a test of $H_0 : \boldsymbol{\beta} = \mathbf{0}$ can be carried out. For testing rare variants it is common to replace the term $\boldsymbol{\beta}' \mathbf{X}_i$ in (7.1) with βr_i , where $r_i = \sum_{j=1}^J X_{ij}$ is the total number of rare variants (e.g. Morris and Zeggini (2010); Yilmaz and Bull (2011)), but this corresponds to using a linear statistic in previous sections and is often ineffective. We consider the case where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_J)'$, in order to examine settings for which causal SNPs may be either deleterious or beneficial. In that case consideration of the power of alternative tests in large samples parallels the discussion in Section 4, as follows.

Let $\hat{\boldsymbol{\beta}}$ be the estimator of $\boldsymbol{\beta}$ based on the model in question and assume that under $H_0 : \boldsymbol{\beta} = \mathbf{0}$, the asymptotic distribution of $\sqrt{n}\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\mathbf{0}$ and covariate matrix $\boldsymbol{\Sigma}$. Following Li and Lagakos (2006), we consider a sequence of contiguous alternatives

$$H_1^{(n)} : \boldsymbol{\beta} = \mathbf{b}/\sqrt{n} \quad (7.2)$$

where $\mathbf{b} = (b_1, \dots, b_J)'$ is a specified vector. Under this sequence as $n \rightarrow \infty$ the distribution of $\sqrt{n}\hat{\boldsymbol{\beta}}$ approaches a multivariate normal distribution with mean \mathbf{b} and covariance matrix $\boldsymbol{\Sigma}$. Thus, asymptotic power for a test statistic can be computed in the same way as in Section 4. Li and Lagakos (2006) compare the “omnibus” test statistic $W = \hat{\boldsymbol{\beta}}' \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\Sigma}}$ is a consistent estimate of $\boldsymbol{\Sigma}$ under H_0 , with linear statistics $Z = \mathbf{a}' \hat{\boldsymbol{\beta}}$. These are analogous to (2.5) and (2.4), respectively. In fact, note that if we consider the linear regression model (4.1) with the ϵ_i independent $N(0, \sigma^2)$ random variables, then $\hat{\boldsymbol{\beta}} = (\tilde{X}' \tilde{X})^{-1} \tilde{X}' \mathbf{Y}$, where $\mathbf{Y} = (Y_1, \dots, Y_n)'$ and \tilde{X} is a centered $n \times J$ matrix whose i 'th row is $\mathbf{X}_i = (X_{i1}, \dots, X_{iJ})' - (\bar{X}_1, \dots, \bar{X}_J)$, and so

$$W = \hat{\boldsymbol{\beta}}' \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\beta}} = \mathbf{S}' \boldsymbol{\Sigma}_S^{-1} \mathbf{S},$$

where \mathbf{S} has j 'th element $S_j = \sum_{i=1}^n (X_{ij} - \bar{X}_j) Y_i = \sum_{i=1}^n X_{ij} (Y_i - \bar{Y})$.

A similar results hold in the logistic model (7.1). The “omnibus” Wald statistic $W^* = \hat{\boldsymbol{\beta}}' \widehat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\beta}}$ is asymptotically equivalent to the likelihood score statistic W_S for testing $\boldsymbol{\beta} = \mathbf{0}$ and when there are no covariates \mathbf{v}_i , W_S is exactly $W_H = \mathbf{S}' \boldsymbol{\Sigma}_S^{-1} \mathbf{S}$. For the case where covariates \mathbf{v}_i are present, an asymptotically equivalent statistic to W^* is based on the likelihood score for prospective sampling under (7.1). The score statistic for testing $\boldsymbol{\beta} = \mathbf{0}$ is easily found as

$$\mathbf{U} = \sum_{i=1}^n (Y_i - \hat{\mu}_i) \mathbf{X}_i, \quad (7.3)$$

where $\hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\boldsymbol{\gamma}}' \mathbf{v}_i} / (1 + e^{\hat{\beta}_0 + \hat{\boldsymbol{\gamma}}' \mathbf{v}_i})$ and $\hat{\beta}_0, \hat{\boldsymbol{\gamma}}$ are estimated from (7.1) when $\boldsymbol{\beta} = \mathbf{0}$. It also follows from standard maximum likelihood large sample theory that the covariance matrix of \mathbf{U} under H_0 is estimated consistently by

$$\widehat{\boldsymbol{\Sigma}}_U = \widehat{Var}(\mathbf{U}) = \left(\sum_{i=1}^n \hat{\sigma}_i^2 \mathbf{X}_i \mathbf{X}_i' \right) - \left(\sum_{i=1}^n \hat{\sigma}_i^2 \mathbf{X}_i \tilde{\mathbf{v}}_i' \right) \left(\sum_{i=1}^n \hat{\sigma}_i^2 \tilde{\mathbf{v}}_i \tilde{\mathbf{v}}_i' \right)^{-1} \left(\sum_{i=1}^n \hat{\sigma}_i^2 \tilde{\mathbf{v}}_i \mathbf{X}_i' \right), \quad (7.4)$$

where $\hat{\sigma}_i^2 = \hat{\mu}_i(1 - \hat{\mu}_i)$ and $\tilde{\mathbf{v}}_i = (1, \mathbf{v}_i')'$. These correspond to results given by Lin and Tang (2011), who consider linear statistics based on linear combinations of the elements U_1, \dots, U_J of \mathbf{U} . The statistic (7.3) and variance estimate (7.4) can be shown to apply under case-control sampling and they give test statistics such as $W_H^* = \mathbf{U}' \widehat{\boldsymbol{\Sigma}}_U^{-1} \mathbf{U}$ and $W_L^* = (\mathbf{w}' \mathbf{U}) / (\mathbf{w}' \widehat{\boldsymbol{\Sigma}}_U^{-1} \mathbf{w})$, which correspond to W_H and W_L in preceding sections. When there are no covariates \mathbf{v}_i , it is readily seen that (7.3) reduces to (3.1) and that (7.4) equals $(n - 1)/n$ times (3.3). It should be noted that when covariates \mathbf{v}_i are present, the normal approximations considered earlier apply, but the permutation distribution p-values do not unless the \mathbf{X}_i are independent of the \mathbf{v}_i . Lin and Tang (2011) suggest a parametric bootstrap as an alternative, based on randomly generating responses from the fitted null model based on $\hat{\beta}_0, \hat{\boldsymbol{\gamma}}$. We note

this can be computationally intensive in case-control settings

It should be mentioned that in the case of quantitative Y-dependent sampling and models with supplementary covariates \mathbf{v}_i as in (7.1), adjustments to estimating functions (e.g. based on inverse probability of selection weights) are needed; this is beyond our present scope.

8 Discussion

In this paper, we have compared tests of association between rare variants and phenotypes within a unified framework which gives theoretical insights about the performance of the methods. Our treatment handles all types of phenotypes and can deal with phenotype-dependent sampling of individuals and correlation among the SNPs under consideration. One of the important conclusions of this work is that depending on the alternative, methods can have greatly varying power. We found in extensive numerical studies that, as expected, when both deleterious and protective SNPs are present the quadratic test statistics are much better. They also outperform linear statistics in settings where causal SNPs are all deleterious (or all protective), but a substantial fraction of the SNPs are neutral (not associated with the phenotype). Although they perform well across a broader range of settings, relative performance can vary according to the fraction of causal SNPs and importantly, according to whether causal affects are associated with lower MAF. Our results also indicate that power to detect moderate levels of association is not high unless sample sizes are very large or a high proportion of the chosen SNPs are causal. Consequently it is critical to obtain relevant biological information that can guide the selection of SNPs or weighting strategy for pooled association testing (e.g. King et al. (2010)).

Our results supplement these of Basu and Pan (2011), and a brief comparison is

useful. They found similar results to ours in simulation studies for case-control scenarios, concerning the performance of linear statistics. Among the quadratic statistics, they found that C-alpha/SSU type statistics $W_C = \mathbf{S}'\mathbf{S}$ was generally best, and superior to the Hotelling statistics (2.5). However, their simulation scenarios did not include cases where causal effects were associated with SNPs having smaller MAFs. Our numerical studies and investigation of GAW17 data indicate the importance of relationships between SNP effects and MAFs. As an approach to rare variant testing in the absence of strong prior information, we support the recommendation of Basu and Pan (2011), to perform tests using both linear and quadratic statistics. In Derkach et al. (2012), we examine statistics that combine evidence from linear and quadratic tests.

Acknowledgements

This study was supported by an Ontario Graduate Scholarship (OGS) and CIHR Strategic Training for Advanced Genetic Epidemiology (STAGE) fellowship to A. Derkach, and Natural Sciences and Engineering Research Council of Canada (NSERC) grants to J. Lawless and L. Sun. The authors would like to thank the Genetic Analysis Workshop (GAW) committee and the 1000 Genomes Project for providing the GAW17 application data and Dr. Andrew Paterson for constructive discussions.

References

- 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.
- Almasy, L., Dyer, T., Peralta, J., Kent, J., Charlesworth, J., Curran, J., and Blangero, J. (2011). Genetic analysis workshop 17 mini-exome simulation. *BMC Proceedings*, 5(Suppl 9):S2.
- Asimit, J. and Zeggini, E. (2010). Rare variant association analysis methods for complex traits. *Annual Review of Genetics*, 44(August):293–308.
- Bansal, V., Libiger, O., Torkamani, A., and Schork, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet*, 11(11):773–85.
- Basu, S. and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, pages n/a–n/a.
- Derkach, A., Lawless, J., and Sun, L. (2012). Combining p-values from linear and quadratic tests for rare variants provides robust statistics across genetic models. Presented at Canadian Human and Statistical Genetics Meeting, Niagara on the Lake, ON.
- Duchesne, P. and de Micheaux, P. L. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. *Computational Statistics and Data Analysis*, 54:858–862.
- Goeman, J. J., Van De Geer, S. A., and Van Houwelingen, H. C. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):477–493.
- Han, F. and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*, 70(1):42–54.
- Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367.
- Hoffmann, T. J., Marini, N. J., and Witte, J. S. (2010). Comprehensive approach to analyzing rare genetic variants. *PLoS ONE*, 5(11):e13584.
- Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, 2 edition.
- King, C. R., Rathouz, P. J., and Nicolae, D. L. (2010). An evolutionary framework for association testing in resequencing studies. *PLoS Genet*, 6(11):e1001202.
- Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics in press*.

- Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83(3):311–321.
- Li, Q. H. and Lagakos, S. W. (2006). On the relationship between directional and omnibus statistical tests. *Scandinavian Journal of Statistics*, 33(2):239–246.
- Lin, D.-Y. and Tang, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet.*, 89(3):354 – 367.
- Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384.
- Manolio, T. A., Brooks, L. D., and Collins, F. S. (2008). A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*, 118(5):1590–605.
- Manolio, T. A. and Collins, F. S. (2009). The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy. *Annu Rev Med*, 60(1):443–456.
- Manolio, T. A., Collins, F. S., Cox, N. J., and et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.
- Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Academic Press.
- Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res*, 615(1-2):28–56.
- Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*, 34(2):188–193.
- Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., and et al. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322.
- Pan, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497–507.
- Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., and et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.*, 86(6):832–838.
- Rao, C. (1973). *Linear statistical inference and its applications*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley.
- Wu, M. C., Lee, S., Cai, T., Li, Y., and et al. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *Am J Hum Genet*.
- Yi, N. and Zhi, D. (2011). Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol*, 35(1):57–69.

Yilmaz, Y. and Bull, S. (2011). Are quantitative trait-dependent sampling designs cost effective for analysis of rare and common variants? *BMC*, Proc 5 (suppl 9).

Figures and Tables

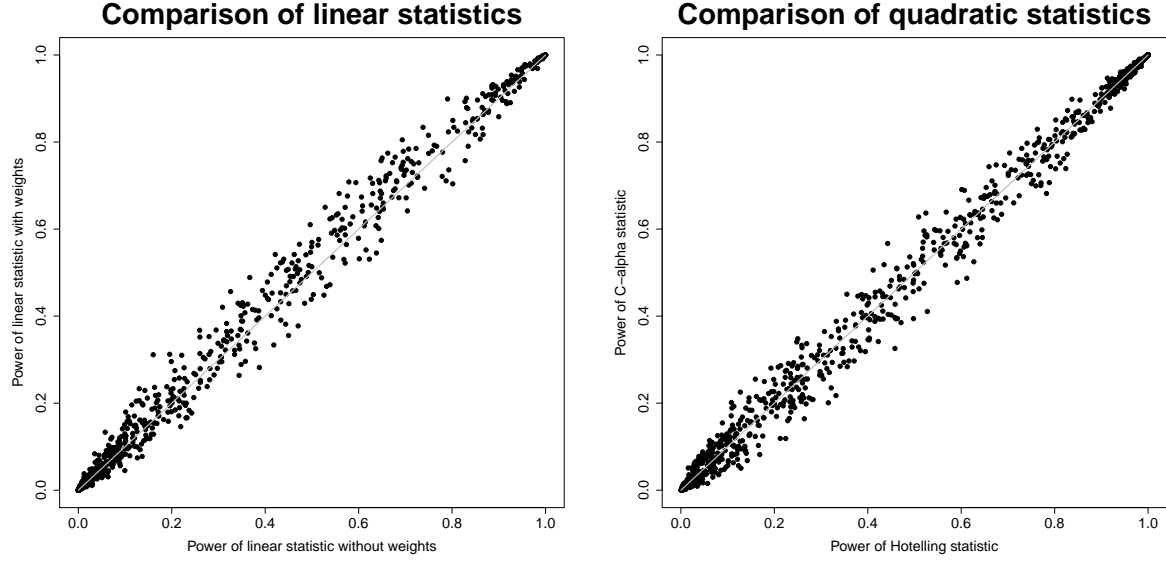


Figure 1:

Comparison of power of linear statistics: linear statistic W_{LS} is given by (2.4) with w_j set to 1 and linear statistic W_{LW} given by (2.4) with w_j set to $1/\sqrt{p_j(1-p_j)}$.

Comparison of power of quadratic statistics: Hotelling statistic is W_H (2.5) and C-alpha statistic is W_Q (2.6) with matrix A set to the identity I .

The points in each plot show the power of each statistic for each of the 1000 randomly chosen models with the genetic effect of a SNP independent of its MAF.

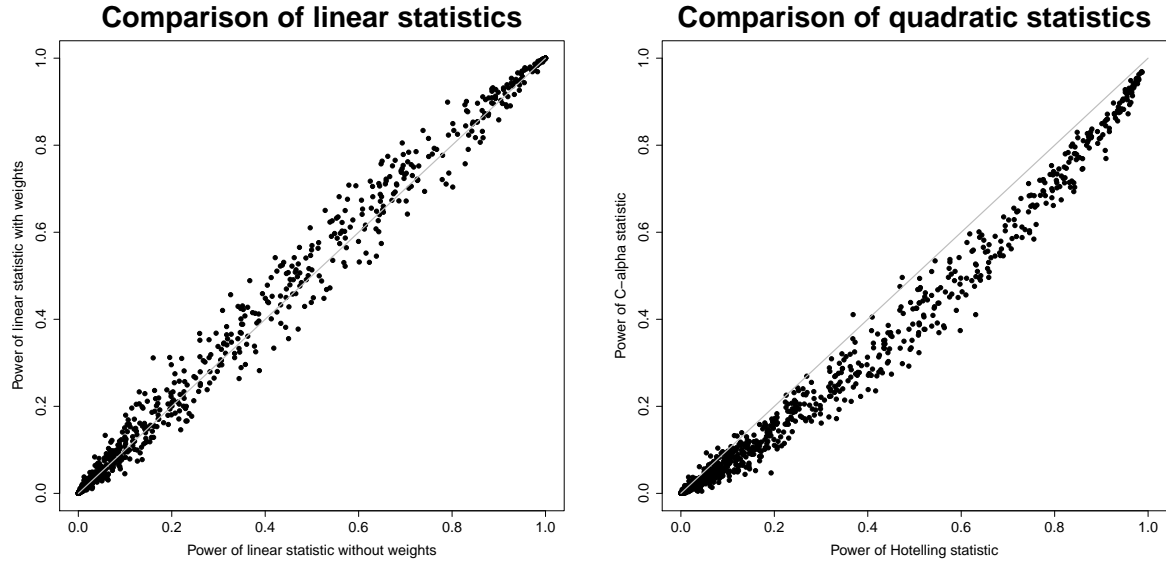


Figure 2:

Comparison of power of linear statistics: linear statistic W_{LS} is given by (2.4) with w_j set to 1 and linear statistic W_{LW} given by (2.4) with w_j set to $1/\sqrt{p_j(1-p_j)}$.

Comparison of power of quadratic statistics: Hotelling statistic is W_H (2.5) and C-alpha statistic is W_Q (2.6) with matrix A set to the identity I .

The points in each plot show the power of each statistic for each of the 1000 randomly chosen models with with explained variation of a single SNP independent of its MAF.

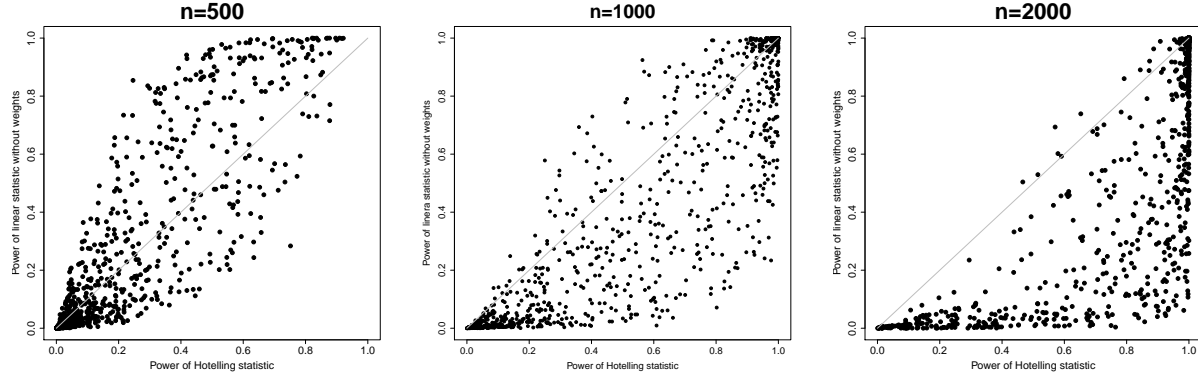


Figure 3:

Comparison of power of linear and quadratic statistics: W_L (2.4) with weights w_j set to 1 (W_{LS}) and the Hotelling statistic W_H (2.5)

Points represent the power for each of the 1000 randomly chosen models with parameters described in Section 4 and with genetic effect of a SNP independent of its MAF. n - sample size.

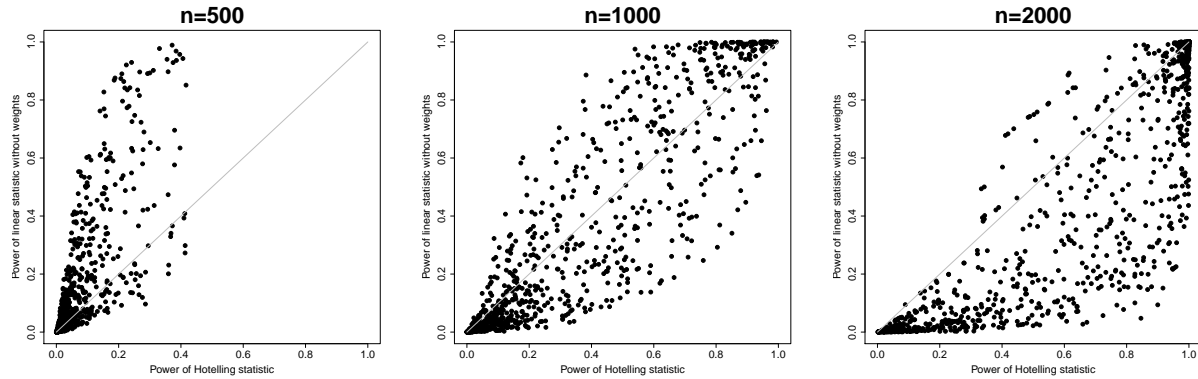
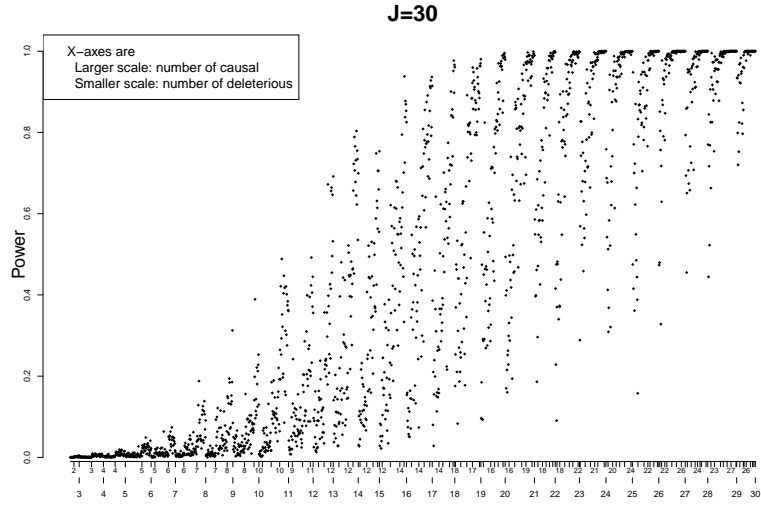


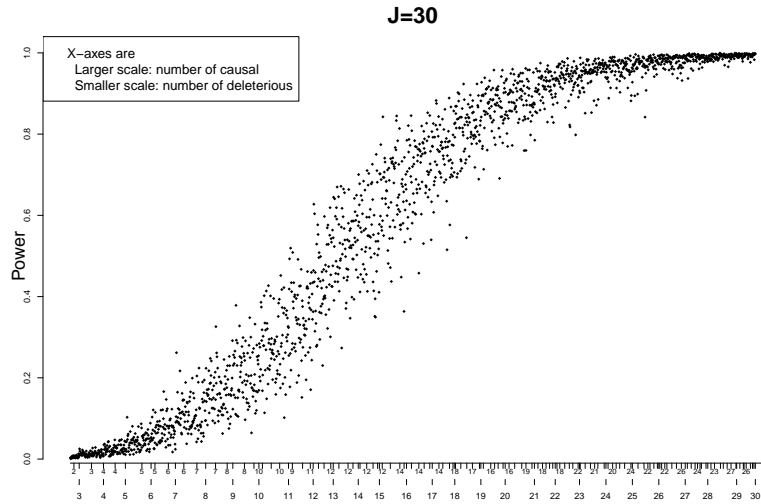
Figure 4:

Comparison of power of linear and quadratic statistics: W_L (2.4) with weights w_j set to 1 (W_{LS}) and the Hotelling statistic W_H (2.5)

Points represent the power for each of the 1000 randomly chosen models with parameters described in Section 4 and with explained variation of a single SNP independent of its MAF. n - sample size.



(a)



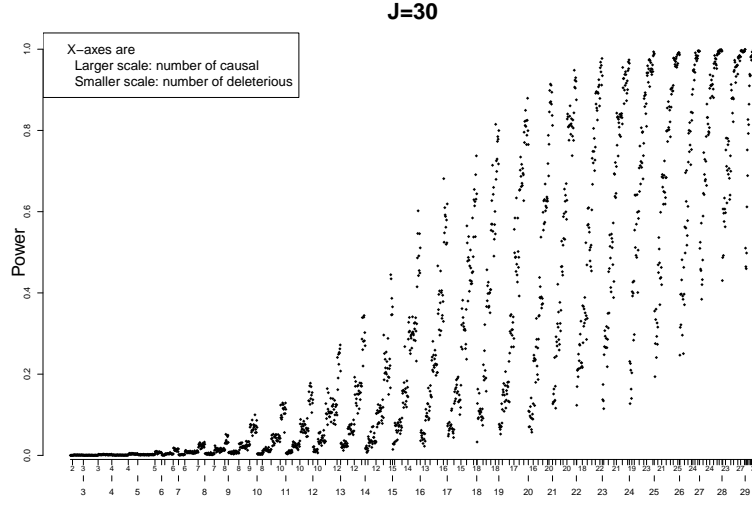
(b)

Figure 5: Empirical power of the statistics represented as function of two parameters: J -number of SNPs, $n = 1000$

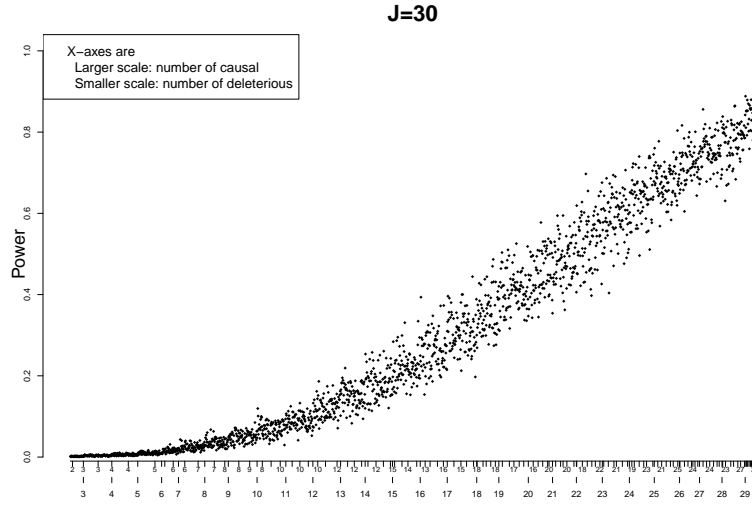
On the X-axis: number of causal SNPs, larger scale; number of deleterious SNPs, smaller scale

(a) Linear statistic, (b) Hotelling statistic.

Results are based on 10000 randomly chosen models with parameters described in Section 4, with the genetic effects of SNPs independent of MAF.



(a)



(b)

Figure 6: Empirical power of the statistics represented as function of two parameters: J -number of SNPs, $n = 1000$

On the X-axis: number of causal SNPs, larger scale; number of deleterious SNPs, smaller scale

(a) Linear statistic, (b) Hotelling statistic.

Results are based on 10000 randomly chosen models with parameters described in Section 4, with explained variation of single SNPs independent of MAF.

$\alpha = 0.05$			$\alpha = 0.01$		
Methods			Methods		
J	LS	QS(Chsq)	LS	QS(Chsq)	
10	0.049	0.041	0.0098	0.006	
20	0.050	0.041	0.0099	0.006	
50	0.050	0.041	0.0099	0.007	
100	0.050	0.041	0.0100	0.007	

$\alpha = 1 \cdot 10^{-3}$			$\alpha = 1 \cdot 10^{-4}$		
Methods			Methods		
J	LS	QS(Chsq)	LF	QS(Chsq)	
10	$0.94 \cdot 10^{-3}$	$0.3 \cdot 10^{-3}$	$0.84 \cdot 10^{-4}$	$0.14 \cdot 10^{-4}$	
20	$0.99 \cdot 10^{-3}$	$0.4 \cdot 10^{-3}$	$0.88 \cdot 10^{-4}$	$0.15 \cdot 10^{-4}$	
50	$1.01 \cdot 10^{-3}$	$0.4 \cdot 10^{-3}$	$1.02 \cdot 10^{-4}$	$0.38 \cdot 10^{-4}$	
100	$1.01 \cdot 10^{-3}$	$0.5 \cdot 10^{-3}$	$0.99 \cdot 10^{-4}$	$0.48 \cdot 10^{-4}$	

Table 1: Empirical type 1 errors for two tests of association with a binary outcome. LS: normal approximation for the linear statistic (2.4); QS(Chsq): chi-square approximation for the Hotelling quadratic statistic (2.5)

Sample size is 500 cases and 500 controls; MAF of each SNP is 0.5%; the number of simulation replicates is 10^6 .

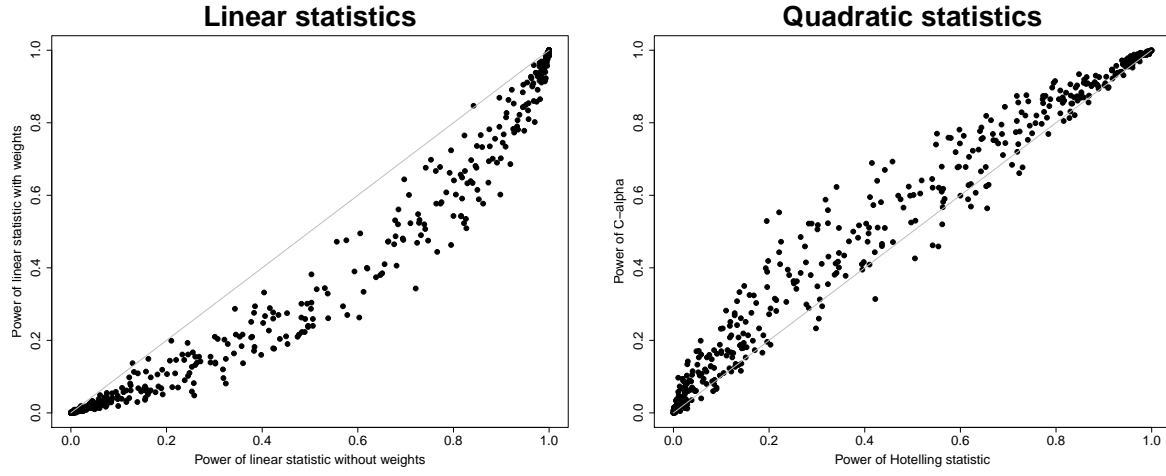


Figure 7:

Comparison of power between linear statistics: (2.4) with weights w_j set to 1 (W_{LS}) and with weights w_j set to $1/\sqrt{p_j(1-p_j)}$ (W_{LW}).

Comparison of power between quadratic statistics: Hotelling statistic W_H (2.5) and C-alpha statistic W_Q (2.6) with matrix A set to the identity I .

Points represent the power for each of the 1000 randomly chosen models with parameters as described in Section 5, with the genetic effect of a SNP β_j independent of its MAF p_j .

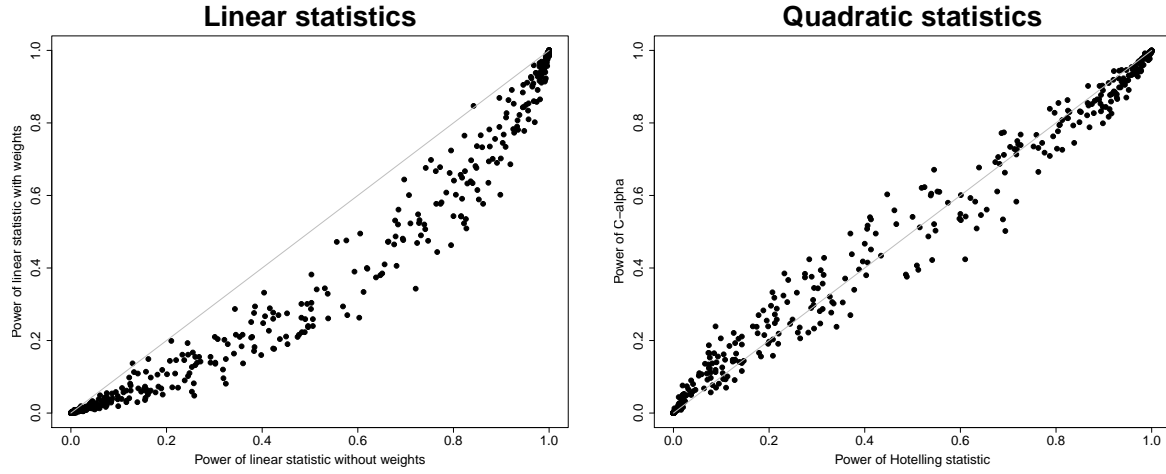


Figure 8:

Comparison of power between linear statistics: (2.4) with weights w_j set to 1 (W_{LS}) and with weights w_j set to $1/\sqrt{p_j(1-p_j)}$ (W_{LW}).

Comparison of power between quadratic statistics: Hotelling statistic W_H (2.5) and C-alpha statistic W_Q (2.6) with matrix A set to the identity I .

Points represent the power for each of the 1000 randomly chosen models with parameters as described in Section 5, with $e^{|\beta_j|} = C/\sqrt{p_j(1-p_j)}$.

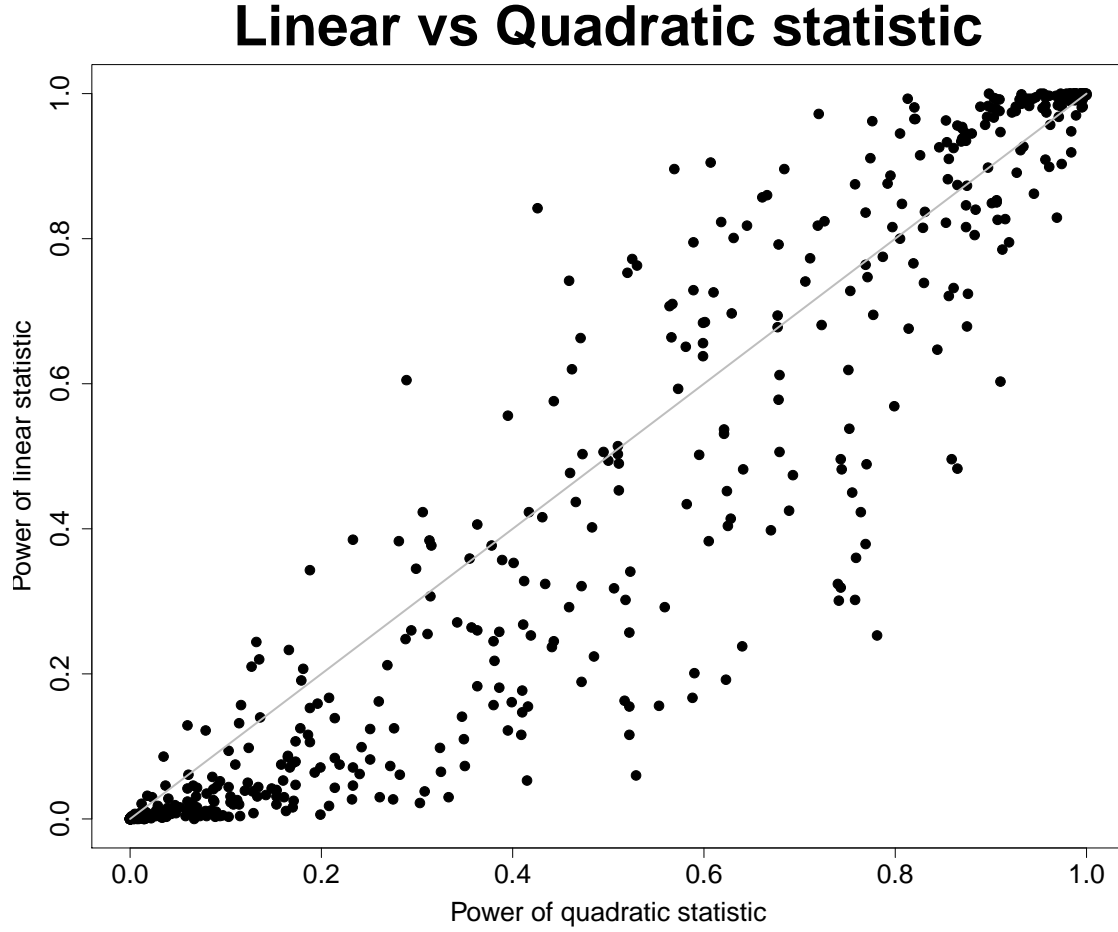


Figure 9:

Comparison of power between linear statistic (2.4) with weights w_j set to 1 (W_{LS}) and C-alpha statistic W_Q (2.6) with matrix A set to the identity I .

Points represent the power for each of the 1000 randomly chosen models with parameters as described in Section 5, with the genetic effect of a SNP β_j independent of its MAF p_j .

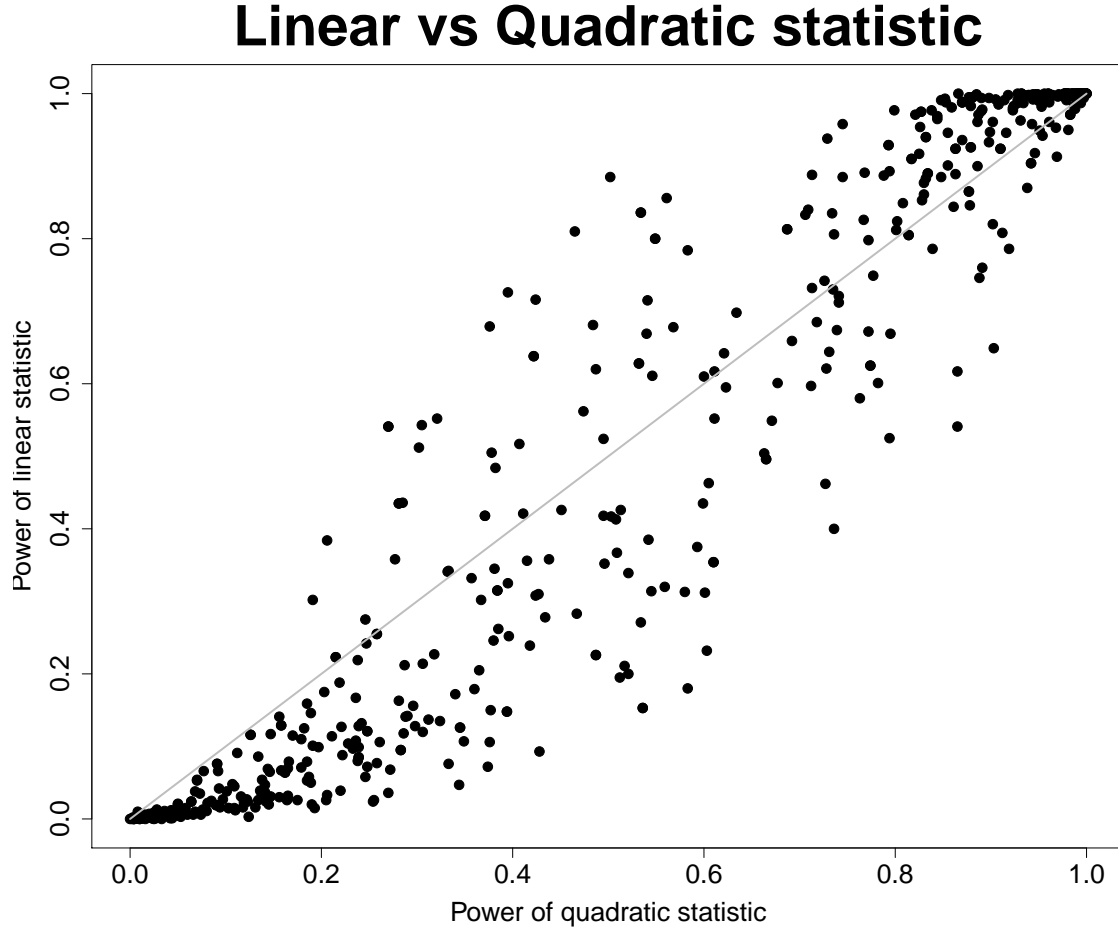
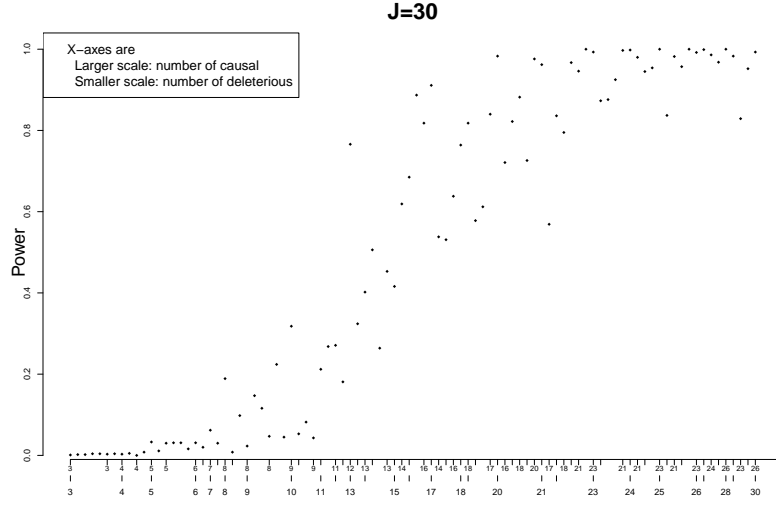


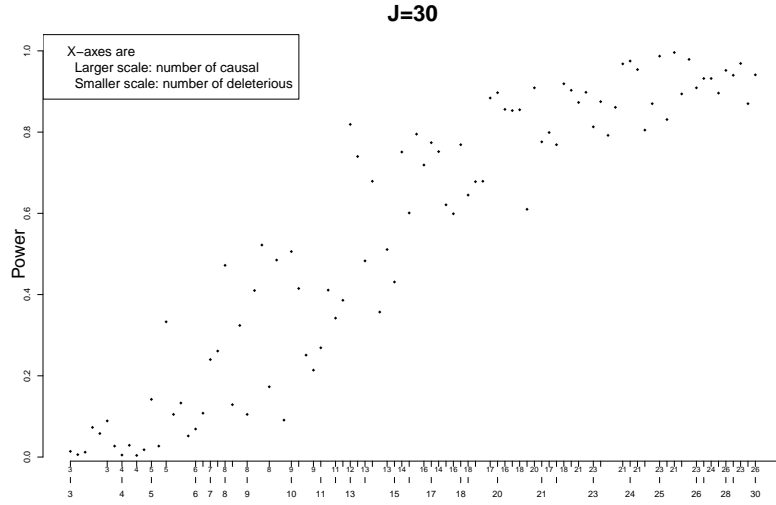
Figure 10:

Comparison of power between linear statistic (2.4) with weights w_j set to 1 (W_{LS}) and C-alpha statistic W_Q (2.6) with matrix A set to the identity I .

Points represent the power for each of the 1000 randomly chosen models with parameters as described in Section 5, with parameters described in Section 5 and with with $e^{|\beta_j|} = C/\sqrt{p_j(1-p_j)}$.



(a)



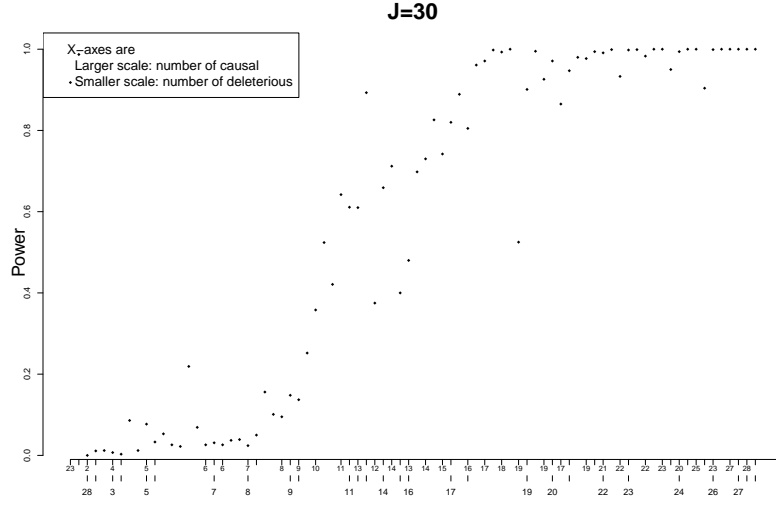
(b)

Figure 11: Empirical power of the statistics represented as function of two parameters: J -number of SNPs, $n = 1000$

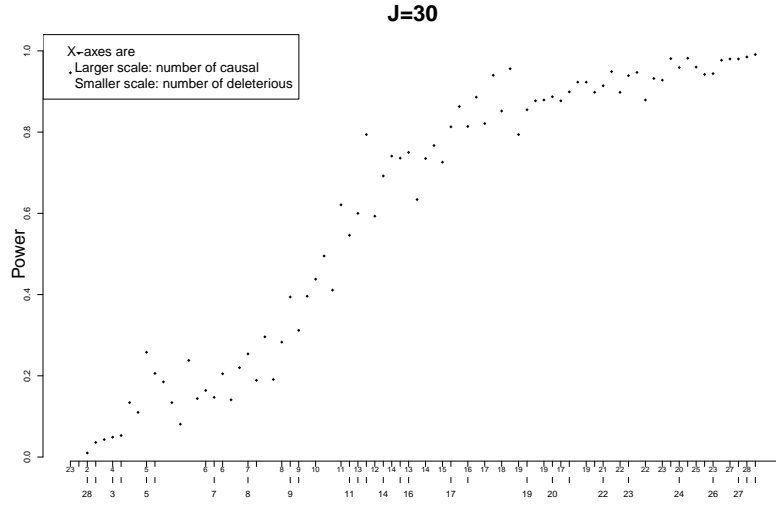
On the X-axis: number of causal SNPs, larger scale; number of deleterious SNPs, smaller scale

(a) Linear statistic, (b) Hotelling statistic.

Results are based on 500 randomly chosen models with parameters described in Section 5, with the genetic effect of a SNP β_j independent of its MAF p_j .



(a)



(b)

Figure 12: Empirical power of the statistics represented as function of two parameters: J -number of SNPs, $n = 1000$

On the X-axis: number of causal SNPs, larger scale; number of deleterious SNPs, smaller scale

(a) Linear statistic, (b) Hotelling statistic.

Results are based on 500 randomly chosen models with parameters described in Section 5 and with $e^{|\beta_j|} = C/\sqrt{p_j(1-p_j)}$.

Gene	SNP Distribution		Ave. MAF of		Avg. Effect of		Power			
	J_C	J_N	J_C	J_N	J_C		Linear W_{LW}	Linear W_{LS}	Quadratic W_C	Quadratic W_H
SIRT1	4, 7		0.27%, 0.22%		0.71		0.44	0.40	0.26	0.39
BCHE	5, 10		0.20%, 0.19%		0.72		0.29	0.35	0.43	0.39
PDGFD	3, 6		0.78%, 0.65%		0.74		0.29	0.43	0.45	0.35
SREBF1	4, 5		0.39%, 0.40%		0.52		0.29	0.27	0.11	0.15
VLDLR	4, 6		0.19%, 1.64%		0.75		0.12	0.08	0.06	0.09
PLAT	4, 7		0.39%, 0.49%		0.68		0.13	0.13	0.06	0.13
RARB	1, 5		0.78%, 0.90%		0.64		0.06	0.025	0.065	0.14
INSIG1	3, 1		0.16%, 3.42%		0.20		0.06	0.06	0.04	0.025
VNN3	2, 2		0.16%, 2.57%		0.37		0.03	0.10	0.06	0.04
LPL	1, 4		0.16%, 0.23%		0.73		0.015	0.03	0.06	0.05
VWF	1, 3		0.16%, 1.90%		0.34		0.02	0.01	0.03	0.01
GCKR	1, 0		1.21%, NA		0.38		0.25	0.25	0.25	0.25
VNN1	0, 3		NA, 0.31%		NA		0.02	0.02	0.04	0.05

Table 2: Power of the four test statistics applied to GAW17 data provided by the 1000 Genomes Project. The four statistics are linear W_{LS} and W_{LW} , quadratic W_C and W_H . All causal variants were designed by GAW17 to have the same direction of effect (i.e. minor allele was associated with higher Q2 value). The average genetic effect is the average of β among the causal variants. Power is estimated using all the 200 replicates at the $\alpha = 0.05$ level.

Supplementary Materials

Derivations

Proof of (2.3) with discrete Y

In this case, let Y_1^*, \dots, Y_k^* denote the K distinct values of Y_i ($i = 1, \dots, n$), and let n_r^* be the number of values Y_r^* , with $\sum_{r=1}^k n_r^* = n$. Let α_r^* be a rank score assigned to Y_r^* ($r = 1, \dots, k$), centered so that $\sum_{r=1}^k n_r^* \alpha_r^* = 0$. Then under H_0 , we have

$$\Pr(\alpha_i = \alpha_r^*) = \frac{n_r^*}{n},$$

$$\Pr(\alpha_i = \alpha_r^*, \alpha_{i'} = \alpha_s^*) = \frac{n_r^* [n_s^* - I(r = s)]}{n(n-1)}, \quad i \neq i'.$$

Thus $E(\alpha_i) = 0$ and for $i \neq i'$,

$$\begin{aligned} E(\alpha_i \alpha_{i'}) &= \sum_{r=1}^k \sum_{s=1}^k \alpha_r^* \alpha_s^* \Pr(\alpha_i = \alpha_r^*, \alpha_{i'} = \alpha_s^*) \\ &= \sum_{r=1}^k \frac{\alpha_r^{*2} n_r^* (n_r^* - 1)}{n(n-1)} + \sum_{r \neq s} \alpha_r^* \alpha_s^* \frac{n_r^* n_s^*}{n(n-1)} \\ &= \sum_{r=1}^k \sum_{s=1}^k \frac{\alpha_r^* \alpha_s^* n_r^* n_s^*}{n(n-1)} - \sum_{r=1}^k \frac{\alpha_r^{*2} n_r^*}{n(n-1)} \\ &= - \sum_{i=1}^n \frac{\alpha_i^2}{n(n-1)}. \end{aligned}$$

Thus $E_P(S_j) = 0$ and

$$\begin{aligned}
E_P(S_j S_\ell) &= \sum_{i=1}^n \sum_{i'=1}^n E(\alpha_i \alpha_{i'}) X_{ij} X_{i'\ell} \\
&= \sum_{i=1}^n E(\alpha_i^2) X_{ij} X_{i\ell} + \sum_{i \neq i'} \sum_{i'=1}^n E(\alpha_i \alpha_{i'}) X_{ij} X_{i'\ell} \\
&= \sum_{i=1}^n \left(\sum_{r=1}^n \frac{\alpha_r^2}{n} \right) X_{ij} X_{i\ell} + \sum_{i \neq i'} \sum_{i'=1}^n \left(- \sum_{r=1}^n \frac{\alpha_r^2}{n(n-1)} \right) X_{ij} X_{i'\ell},
\end{aligned}$$

which reduces to (2.3).

Proof of (3.3)

Since $E_P(S_j) = 0$ the covariance of S_j and S_ℓ is

$$E_P(S_j S_\ell) = \sum_{i=1}^n \sum_{i'=1}^n E_P(Y_i Y_{i'}) (X_{ij} - \bar{X}_j) (X_{i'\ell} - \bar{X}_\ell),$$

using the fact that (3.1) also equals $\sum_{i=1}^n (X_{ij} - \bar{X}_j) Y_i$. Thus

$$\begin{aligned}
E_P(S_j S_\ell) &= \sum_{i=1}^n E_P(Y_i^2) (X_{ij} - \bar{X}_j) (X_{i\ell} - \bar{X}_\ell) \\
&\quad + \sum_{i \neq i'} \sum_{i'=1}^n E_P(Y_i Y_{i'}) (X_{ij} - \bar{X}_j) (X_{i'\ell} - \bar{X}_\ell) \\
&= \left(\frac{n_1}{n} \right) \sum_{i=1}^n (X_{ij} - \bar{X}_j) (X_{i\ell} - \bar{X}_\ell) \\
&\quad + \frac{n_1(n_1-1)}{n(n-1)} \sum_{i \neq i'} \sum_{i'=1}^n (X_{ij} - \bar{X}_j) (X_{i'\ell} - \bar{X}_\ell) \\
&= \left(\frac{n_1}{n} \right) \left(1 - \frac{(n_1-1)}{n-1} \right) \sum_{i=1}^n (X_{ij} - \bar{X}_j) (X_{i\ell} - \bar{X}_\ell).
\end{aligned}$$

Noting that $\sum_{i=1}^n X_{ij} = m_j$ and that $\sum_{i=1}^n X_{ij} X_{i\ell} = m_{j\ell}$, we obtain (3.5); noting that

$\sum_{i=1}^n X_{ij}^2 = m_j$, we also obtain (3.3).